DOCUMENT RESUME

ED 161 884 TM 007 171

AUTHOR Patience, Wayne M.; Reckase, Mark D.
TITLE Self-Paced Versus Paced Evaluation Utilizing Computerized Tailored Testing.
PUB DATE Mar 78
CONTRACT N00014-77-C-0097
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Tests; Flexible Scheduling; Higher Education; *Pacing; *Testing
IDENTIFIERS Computer Assisted Testing; *Paper and Pencil Tests; *Tailored Testing

ABSTRACT

The feasibility of implementing self-paced computerized tailored testing evaluation methods in an undergraduate measurement and evaluation course, and possible differences in achievement levels under a paced versus self-paced testing schedule were investigated. A maximum likelihood tailored testing procedure based on the simple logistic model had previously been used for evaluation in this course; however, scheduling of the testing sessions had been determined by the instructor. The basic thrust of the initial question addressed the possibilities of having students determine when they would prefer to take the exams. The study also investigated whether or not there would be significant differences in achievement level of students allowed to schedule their exams and those whose exams were scheduled by the instructor. One hundred and seventy-two undergraduate students participated in the study. Students were randomly assigned to nine experimental groups consisting of combinations of two exams with the following testing schedules: paced tailored test, self-paced tailored test, and traditional paper and pencil test. Results on a comprehensive final were used as dependent measures. Since computerized tailored testing did not, in itself, affect achievement, and since it provides immediate feedback to the student, it is concluded to be an increasingly feasible method of testing. (Author/CTM)

Self-Paced Versus Paced Evaluation Utilizing Computerized Tailored Testing

by

Wayne M. Patience and Mark D. Reckase

University of Missouri-Columbia

Self-Paced Versus Paced Evaluation Utilizing Computerized Tailored Testing

by

Wayne M. Patience and Mark D. Reckase

University of Missouri-Columbia

Abstract

The research investigated the implementation of computerized tailored testing for the measurement of achievement under paced versus self-paced examination conditions. One hundred and seventy-two undergraduate students in an introductory measurement and evaluation course participated in the study. Students were randomly assigned to nine experimental groups consisting of combinations of two exams with the following testing schedules: paced tailored test, self-paced tailored test, and traditional paper and pencil test. Results on a comprehensive final were used as dependent measures. The tailored testing procedure was based on the simple logistic model. Attitudinal data was also incorporated in analyses.

Self-Paced Versus Paced Evaluation Utilizing Computerized Tailored Testing

by

Wayne M. Patience and Mark D. Reckase

University of Missouri-Columbia

Objectives of the Inquiry

The two primary objectives of the study described herein were 1) to determine the feasibility of implementing self-paced computerized tailored testing evaluation methods in an undergraduate measurement and evaluation course, and 2) to investigate possible differences in achievement levels under a paced versus self-paced testing schedule. A maximum likelihood tailored testing procedure based on the simple logistic model had previously been used for evaluation in this course, however, scheduling of the testing sessions had been determined by the instructor. The basic thrust of the initial question addressed the possibilities of having students determine when they would prefer to take the exams. Availability of alternate forms is dramatically increased in as much as tailored testing will usually not administer exactly the same test twice. The second question to be investigated was whether or not there would be significant differences in achievement level of students allowed to schedule their exams and those whose exams were scheduled by the instructor.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, 1978. This research was supported by Contract Number N00014-77-C-0097 from the Personnel and Training Research Programs of the Office of Naval Research and a University of Missouri Research Council grant. Mark D. Reckase was principal investigator for both grants.

Secondary objectives included an investigation of two additional questions. Are there differences in achievement levels of students taking paper-pencil tests and those taking exams via computerized tailored testing? Do differences exist in student attitudes toward paper-pencil tests, paced tailored tests, and self-paced tailored tests? A four item Likert type attitude questionnaire was given to determine student attitudes toward the testing procedures. A comprehensive final which all students took at the same time under traditional paper and pencil conditions assessed the overall achievement level for each student.

Instrumentation

All items administered on both the paper-pencil tests and computerized tests were of the multiple-choice variety. The items administered on the tailored tests were calibrated using the Rasch simple logistic model, and stored in an item pool to be accessed by the procedure. The methods employed for item selection and ability estimation by the computerized tests relate the probability of a correct response to the ability of the person and the easiness of the item. Item pools were constructed of items determined to be of sufficient quality and content across the continuum of easiness. The item calibration derived from the simple logistic model yields one parameter, easiness, for each item. When an examinee is tested initially, the first item administered has a probability of .5 of a correct response for a person of average ability. If a correct response is obtained, the next item selected is more difficult. If the examinees response is incorrect, an easier item is administered. When both a correct and incorrect response has been obtained, the maximum-likelihood procedure estimates ability using an iterative search for the mode of the likelihood distribution. The tailored test continues the cycle of selecting and administering

items, recording the response pattern, and making ability estimates, until the item pool has been depleted of appropriate items for the examinee's estimated ability, ability had been estimated with sufficient accuracy, or twenty items have been administered. For a more complete description of the tailored testing procedure, see Lord, 1970; Weiss, 1974; Reckase, 1974; or Patience, 1977. This procedure has been demonstrated to have comparable reliability with traditional paper and pencil tests which have many more items administered thus requiring much more time to administer (Reckase, Note 1). Also, test security is much less of a problem due to the previously cited readily available alternate forms.

The computer used in administering the tailored tests was an IBM 370/168 with time sharing capability when linked with remote terminals via phone lines. The terminal used for display of the test items and recording of examinees response patterns was a Beehive Mini-Bee II cathode ray terminal.

Methods

One hundred and seventy-two undergraduate students in an introductory course in measurement and evaluation participated in the study. Students were randomly assigned to the experimental groups which consisted of the nine possible combinations of two exams with the following testing conditions: paced tailored test, self-paced tailored test, and traditional paper-and-pencil test. This pairing of exams with the three testing modalities provided the basis for studying the feasilility of implementing student self pacing of their examinations. The students randomly assigned to the nine experimental groups consisted of those students who volunteered for the study. Students that did not volunteer for the experiment were also incorporated into the analyses as a "non experimental" external control group. Results on a comprehensive final, which all students

took in the traditional manner, were used as dependent measures along with the students' total score in the class.

Depending upon the experimental group in which the student was randomly assigned, he or she took the first two exams in the course in one of the following conditions: exam one self-paced and exam two self-paced (SPSP), exam one self-paced and exam two paced (SPP), exam one self-paced and exam two traditional (SPT), exam one paced and exam two self-paced (PSP), exam one paced and exam two paced(PP), exam one paced and exam two traditional (PT), exam one traditional and exam two self-paced (TSP), exam one traditional and exam two paced (TP), and exam one and two both traditional (TT). The TT group and the non-experimental external control group (EC) were compared to determine whether differences existed between those who volunteered and those who did not volunteer. Students were informed via a handout with their name on it how they were to take the first two exams in the course. They were so acquainted with the procedure they were to follow depending upon how their exams were to be administered. If an exam was to be taken traditionally, the date was specified and they took the fifty item multiple-choice test in a group. If an exam was scheduled as paced, they were told to come in during a period amenable for them but within a specified time frame of a few days. If an exam was to be taken self-paced, the student was informed that he could come in to the tailored testing laboratory and schedule a time at which he or she would like to take that particular exam. Under the self-paced condition, students were permitted to take the exam as many times as they cared to until they were satisfied with the grade that they had achieved. Therefore, as was pointed out in the individualized instruction handout, a student could feasibly take a given exam even before instruction in the course had completed that unit. If they scored well on the tailored test over this material,

as would be the case if a student was well versed in the material from past training and experience, they would most likely forgo attending the class during this particular set of instruction.

The third exam for everyone was administered under traditional circumstances i.e. paper and pencil, and at the same time in a large group. This comprehensive exam of one hundred items was broken down into three parts. Part one consisted of fifty items over the last one-third of the course. Part two of the exam had twenty-five items covering the first one-third of the course or exam one material, and part three consisted of twenty-five items measuring achievement of the middle one-third or exam two material. The total score on the comprehensive final was also recorded.

Results

The following data was collected on all of the experimental groups. On exams one and two, standard scores (Z) were recorded if the examinee took the traditional multiple-choice fifty item paper and pencil test. If the test was taken on the computer terminal under paced or self-paced conditions, log ability scores were recorded. Standard scores were recorded for each of the three parts as well as the total on exam three for all experimental groups. The log ability scores were converted to standard scores for the purpose of obtaining a total Z-score which consisted of two times the exam three total plus exam one score plus exam two score for each student in the course. The primary dependent variables utilized for evaluating possible achievement level differences included: 1) the total standard score for exam three taken as a whole (TOTAL), 2) the standard score for part one of exam three which covered the last one-third of the course material (PART 1), 3) the standard score for part two of exam three which was a retention

measure of the first exam material (PART 2). 4) the standard score for part three of exam three which was a retention measure for exam two material (PART 3), and 5) the total score for the course (Total Z). Table 1 presents the cell means for each of these dependent measures for each of the testing conditions for exams one and two as well as the means for the external control group.

Insert Table 1 about here

Of special interest was whether or not significant achievement level differences existed between those students whose exams were scheduled by the instructor as contrasted with achievement level of students allowed to schedule their own exams. Also of concern, was whether differences existed among students' achievement when exams were administered traditionally with paper and pencil as opposed to exams administered via computerized tailored testing. With respect to this latter investigation, careful attention was directed to scaling the log ability estimates obtained from computerized tailored testing to the standard scores resultant from traditional paper and pencil testing. In addition to comparisons of achievement level for self-paced and paced tailored tests versus traditional testing of those students who volunteered and therefore were randomly assigned to experimental treatment conditions, was the comparison of achievement for students who did not volunteer as contrasted with those who did voluntarily participate. The external control group was, therefore, utilized in making a determination as to whether or not a selection effect occurred. The generalizability of results were thereby improved by the inclusion of the external control group into analyses. While research investigating the operating characteristics of computerized tailored

testing has been enhanced by utilization in actual classroom settings, students in previous studies were found to resent arbitrary assignment to experimental groups which were evaluated via computerized tests if they had not been given the opportunity to specify whether or not they were willing to participate in such a study. When grades have been assigned by innovative and unfamiliar methods, students have exhibited concern and apprehension. This may suggest an advantageous factor related to motivation of students when addressing the use of computerized testing in studies where grades were assigned on the basis of these tests as opposed to simulated studies or research in which students participate and received extra credit for merely taking part.

Analyses of variance were performed for each of the respective dependent variables previously delineated. The five analysis of variance tables are presented below in table two for the three by three factorial design with an external control group. The results presented have only three occurences of significant F values. These included: differences among the session one testing (S1) conditions for dependent variable Part 1, and differences among the session two testing (S2) conditions for dependent variables Total and Part 3.

Insert Table 2 about here

Due to the compounding of the alpha error by repeated analyses of variance on the different dependent variables, at least one of the significant findings may be resultant of chance error instead of the existence of a true difference. A venturesome postulate has been suggested by consideration of a contrasting trend. Across exam one conditions, the students tested traditionally tended to score a little better overall, whereas across exam two conditions, the paced computerized test group consistently tended to score higher taken as a whole. Therefore, if one

was to hazard a discounting of one of the significant results, one could suggest that the difference across S1 for dependent variable Part 1 may not reflect a true difference. In terms of S2 conditions for dependent measures Total and Part 3, the results suggested that the paced tailored test group scored better than the traditionally tested group.

The findings, more importantly, supported the null hypothesis that overall differences between self-paced versus paced testing groups did not occur. There also did not appear to be significant overall achievement differences between individuals tested traditionally as opposed to those who were tested by the computerized test. None of the interactions of S1 and S2 for the respective dependent variables were significant. Also, the external control group's performance was not significantly different from the other nine groups.

Aptitude data was collected where available. This consisted of the college grade point average for each of the junior and senior level students in the course. Missouri Placement Test scores, Missouri College Entrance Test scores, SCAT verbal, quantitative and total scores, and high school rank were also obtained when available. These aptitude measures were found not to be highly predictive of any of the dependent variables when analyzed by multiple regression procedures. Also, a high proportion of missing data on these aptitude measures resulted from incomplete University records.

Whenever an exam was administered via tailored test on the computer terminal. the number of items given was recorded. If the student took an exam under self-paced scheduling conditions, the number of times the test was taken until he or she scored at a level that was satisfactory to the student was recorded. Students taking exams under traditional or paced scheduling conditions were allowed to take the exam only once. The mean number of items presented by the computerized tailored

test was 12.6, representing a substantial reduction in number of items administered as well as time required to administer an individual test. Number of items did not have a significant correlation with the dependent variables sighted earlier. This suggests that having been administered fewer items on the computerized tests did not adversely effect students' performance on any of the components of exam three or on total score for the class. With regard to the number of times students took the self-paced exams, the mean number of exams taken by self-paced students was less than two, suggesting that students under self-paced testing schedules did not take advantage of the provision of being able to take exams as many times as they desired in order to improve their scores. The maximum number of times a test was taken under self-paced conditions was four.

Attitudinal data addressing preference of testing modality, i.e. traditional or computerized tailored test, was collected using a four item Likert type attitude questionnaire. The following dimensions were measured: time pressure, perceived difficulty, anxiety, and overall preference. Table 3 presents descriptive statistics in the form of frequency distributions.

Insert Table 3 about here

The totals for frequency of responses reflect some students who did not respond to the attitude items. Overall trends appear to suggest that students found the tailored test to have less time pressure and about as difficult as traditional tests. They were about equally divided as to amount of anxiety associated with the two testing modalities, and for the most part, overall preference was favorable to the computerized test. Attitude measures correlate significantly with one another but not with achievement measures. This has been found to be the

case in other studies performed with this questionnaire and similar students which were tested with the same tailored testing procedure.

Discussion and Conclusions

The investigation of the feasibility of implementing self-paced scheduling of computerized tailored testing found the procedure to be a viable one. There was a tendency for students taking an exam under self-paced scheduling conditions to procrastinate in as much as most students took their exam after the self-paced or traditional group had completed the exam. Although self-paced students were allowed to take the exam as often as they liked, there was not a tendency for them to score higher on overall achievement across the different treatment conditions. There was no evidence that suggested any major discrepancy between achievement level for students taking their exams paper and pencil as opposed to on the computer terminal. Attitude data reflects that students did not find the tailored test to be objectionable on the dimensions measured, and to a large extent would prefer to take their exams on the computer terminal.

One possible suggested account of why senior level students in this particular study did not take full advantage of the self-paced condition was that the course itself was an eight week block class. This possibly did not provide enough time for students, who typically have not been acclimated to self-paced evaluation, to become accustomed to the possibilities provided by the procedure. Further research into this area of the flexibility of computerized tailored testing is needed.

The most important educational implication of this study suggested that computerized tailored testing offers alternative measurement procedures for evaluating pupil achievement without substantial detrimental effects. Computerized tailored testing was found to be a viable method of self-paced evaluation which is

important in as much as educational programs are attempting to adapt to individual differences. This is especially true of computer assisted instruction in which students progress at their own rate, and there is a need for frequent measurement of achievement. Along this line, the computerized test was found to necessitate significantly fewer items and needed less time to administer to each examinee. Ready availability of forms of exams for tailored tests, as a result of its adaptive nature, (Whitely and Dawis, 1974) alleviates burdensome paper work in facilitating the evaluation of students' progress in a given course of instruction.

In as much as computerized tailored testing has been demonstrated not to affect overall achievement in and of itself, the advantage of frequent and immediate feedback to the learner can be gained by use of this type of exam. In short, computerized testing is becoming more and more feasible and study demonstrates it to be a realistic alternative.

Reference Notes

1. Reckase, M. D. The reliability and validity of achievement tests administered using a variable branching tailored testing model. Unpublished manuscript, 1976.

References

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computerized assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Patience, W. M. Description of components in tailored testing. Behavior Research Methods and Instrumentation, 1977, 9(2), 153-157

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 1974, 6, 208-212.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: Psychometric Methods Program, University of Minnesota, December 1974. (NTIS No. AD A004270).

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, (3), 163-178.

Table 1

MEAN TEST SCORES BY TREATMENT CONDITION

Exam 1 Condition	Exam 2 Condition	Dependent Variable	SP	P	T	Combined
		n	10	11	12	33
		Total	53.30	48.36	41.08	47.58
		Part 1	50.80	46.45	40.67	45.97
	SP	Part 2	55.20	52.27	43.58	50.35
		Part 3	53.50	48.91	43.83	48.75
		Total Z	207.80	192.45	176.92	192.39
		n	12	11	9	32
		Total	47.17	54.27	50.56	50.67
		Part 1	47.75	52.36	52.78	50.96
	P	Part 2	47.83	55.27	49.78	50.96
		Part 3	49.50	54.45	50.33	51.43
		Total Z	192.33	213.27	203.11	202.91
		n	11	12	11	34
		Total	50.00	56.08	47.36	51.15
		Part 1	46.64	58.08	48.64	52.12
	T	Part 2	49.18	51.25	49.18	49.87
		Part 3	50.91	54.75	46.18	50.61
		Total Z	199.09	220.25	195.00	204.78
		n	33	34	32	99
		Total	50.16	52.91	46.33	49.83
		Part 1	49.40	52.30	47.36	49.71
	Combined	Part 2	50.74	52.93	47.51	50.41
		Part 3	51.30	52.70	46.78	50.29
		Total Z	199.74	208.66	191.68	200.12

Non-Experimental Control	n	73
	Total	50.05
	Part 1	50.51
	Part 2	44.41
	Part 3	49.67
	Total Z	200.11

Table 2

Analysis of Variance Results

	Source	DF	SS	MS	F
Dependent Variable Total	S1	2	314.46	157.23	1.655
	S2	2	791.76	395.88	4.168*
	S1xS2	4	797.59	199.40	2.099
	Control vs. Others	1	5.38	5.38	.057
	Error	162	15388.88	94.99	

	Source	DF	SS	MS	F
Dependent Variable Part 1	S1	2	799.63	399.82	4.310*
	S2	2	487.10	243.55	2.626
	S1xS2	4	883.94	220.99	2.382
	Control vs. Others	1	35.64	35.64	.384
	Error	162	15027.60	92.76	

	Source	DF	SS	MS	F
Dependent Variable Part 2	S1	2	21.03	10.52	.103
	S2	2	519.20	259.60	2.531
	S1xS2	4	669.78	167.44	1.633
	Control vs. Others	1	31.21	31.21	.304
	Error	162	16613.30	102.55	

	Source	DF	SS	MS	F
Dependent Variable Part 3	S1	2	159.29	79.65	.821
	S2	2	664.27	332.14	3.424*
	S1xS2	4	426.81	106.70	1.100
	Control vs. Others	1	11.83	11.83	.122
	Error	162	15713.71	96.998	

	Source	DF	SS	MS	F
Dependent Variable Total Z	S1	2	3546.94	1773.47	1.534
	S2	2	5326.88	2663.44	2.304
	S1xS2	4	6685.15	1671.29	1.445
	Control vs. Others	1	4.63	4.63	.004
	Error	162	187313.26	1156.25	

* $p < .05$

Table 3

Attitude Items & Response Data

1. Compared to multiple-choice tests, the tailored test has

		Response Frequency	Value* Assigned
(a)	more time pressure.	8	1
(b)	less time pressure.	48	3
(c)	about equal time pressure.	18	2

2. Compared to traditional multiple choice tests, the tailored test is

		Response Frequency	Value Assigned
(a)	easier.	6	3
(b)	harder.	21	1
(c)	about as difficult.	47	2

3. As compared to the traditional multiple-choice test,

		Response Frequency	Value Assigned
(a)	I would rather take the tailored test.	42	3
(b)	I would rather take the traditional test.	22	1
(c)	I prefer both equally well.	10	2

4. Taking the test on the computer makes me

		Response Frequency	Value Assigned
(a)	more anxious than the traditional test.	27	1
(b)	less anxious than a traditional test.	19	3
(c)	about equally as anxious as the traditional test.	28	2

*These values were utilized in coding responses for correlating the items with dependent measures.